A peer-reviewed version of this preprint was published in PeerJ on 3 July 2018.

View the peer-reviewed version (peerj.com/articles/5179), which is the preferred citable publication unless you specifically need to cite this preprint.

Borstein SR, O'Meara BC. 2018. *AnnotationBustR*: an R package to extract subsequences from GenBank annotations. PeerJ 6:e5179 https://doi.org/10.7717/peerj.5179

AnnotationBustR: An R package to extract subsequences from GenBank annotations

Samuel R. Borstein Corresp., [1] , **Brian C. O'Meara** [1]

[1] Department of Ecology & Evolutionary Biology, University of Tennessee - Knoxville, Knoxville, Tennessee, United States

Corresponding Author: Samuel R. Borstein
Email address: sborstei@vols.utk.edu

Background. DNA sequences are pivotal for a wide array of research in biology. Large sequence databases, like GenBank, provide an amazing resource to utilize DNA sequences for large scale analyses. However, many sequences on GenBank contain more than one gene or are portions of genomes, and inconsistencies in the way genes are annotated and the numerous synonyms a single gene may be listed under provide major challenges for extracting large numbers of subsequences for comparative analysis across taxa. At present, there is no easy way to extract portions from multiple GenBank accessions based on annotations where gene names may vary extensively. **Results.** The R package *AnnotationBustR* allows users to extract sequences based on GenBank annotations through the ACNUC retrieval system given search terms of gene synonyms and accession numbers. *AnnotationBustR* extracts portions of interest and then writes them to a FASTA file for users to employ in their research endeavors. **Conclusion.** FASTA files of extracted subsequences and accession tables generated by *AnnotationBustR* allow users to quickly find and extract subsequences from GenBank accessions. These sequences can then be incorporated in various analyses, like the construction of phylogenies to test a wide range of ecological and evolutionary hypotheses.

AnnotationBustR: An R package to extract subsequences from GenBank annotations

Samuel R. Borstein[1] and Brian C. O'Meara[1]

[1]Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN, USA

Corresponding Author:

Samuel R. Borstein[1]

569 Dabney, University of Tennessee, Knoxville, TN 37996-1610, USA

Email address: sborstei@vols.utk.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2920v1 | CC BY 4.0 Open Access | rec: 12 Apr 2017, publ: 12 Apr 2017

Abstract

Background. DNA sequences are pivotal for a wide array of research in biology. Large sequence databases, like GenBank, provide an amazing resource to utilize DNA sequences for large scale analyses. However, many sequences on GenBank contain more than one gene or are portions of genomes, and inconsistencies in the way genes are annotated and the numerous synonyms a single gene may be listed under provide major challenges for extracting large numbers of subsequences for comparative analysis across taxa. At present, there is no easy way to extract portions from multiple GenBank accessions based on annotations where gene names may vary extensively.
Results. The R package *AnnotationBustR* allows users to extract sequences based on GenBank annotations through the ACNUC retrieval system given search terms of gene synonyms and accession numbers. *AnnotationBustR* extracts portions of interest and then writes them to a FASTA file for users to employ in their research endeavors.
Conclusion. FASTA files of extracted subsequences and accession tables generated by *AnnotationBustR* allow users to quickly find and extract subsequences from GenBank accessions. These sequences can then be incorporated in various analyses, like the construction of phylogenies to test a wide range of ecological and evolutionary hypotheses.

Introduction

The use of DNA sequence data is vital for a wide variety of research in evolutionary biology and ecology. Molecular phylogenies, which rely on DNA sequences for their construction, are extremely prevalent in biological research. Whether being used to correct for shared ancestry among organisms (Felsenstein, 1985), or to test hypotheses related phylogeography (Avise et al., 1987), diversification (Hey, 1992; Maddison, 2006), and trait evolution (Bollback, 2006; O'Meara et al., 2006), phylogenies are required. Additionally, the use of phylogenies is important in community ecology to place systems into an evolutionary framework (Webb et al., 2002; Cavender-Bares et al., 2009). The construction of molecular phylogenies for systematic purposes is also a popular tool for taxonomists to identify new taxa and classify organisms (De Queiroz & Gauthier, 1994; Tautz et al., 2003). Some DNA sequences, like the mitochondrial gene cytochrome oxidase subunit I (COI), are also gaining utility as a method to identify and catalog species using DNA barcoding (Hebert et al., 2003; Ratnasingham & Hebert, 2007; Ratnasingham & Hebert, 2013).

Sequence databases like GenBank provide an extremely valuable resource for using DNA sequence data to test evolutionary and ecological hypotheses. With the reduction in cost of DNA sequencing and the advancement of methods to analyze sequence data, the amount of sequence data available for use is growing at a rapid pace. Given that GenBank has over one-trillion sequences from over 370,000 species (Benson et al., 2017) and recent advances in methods to create massive phylogenies using either super-matrix (Driskell et al., 2004; Ciccarelli et al., 2006) or mega-phylogeny approaches (Smith et al., 2009; Izquierdo-Carrasco et al., 2014), many generate large DNA sequence data sets for comparative analyses (Leslie et al., 2012; Rabosky et al., 2013; Spriggs et al., 2014; Zanne et al., 2014; Shi & Rabosky, 2015). Additionally, sequence retrieval within common scripting environments for biological analyses, like R (R Development Core Team, 2017), are made possible with packages like *ape* (Paradis et al., 2004), *rentrez* (Winter, 2016), *reutils* (Schofl, 2015), and *seqinr* (Charif & Lobry, 2007).

While GenBank provides a wealth of sequence data for researchers to use, some of it is rather difficult to manipulate into a useful form. For example, some sequences may be concatenated together, or the only gene sequence available for a species for the locus of interest may be within a mitochondrial or chloroplast genome. Although GenBank's annotation system provides a means to see where a locus of interest is in a genome or concatenated sequence and provides the ability to download it manually, this is extremely time consuming when many accessions are involved and not a feasible way to extract mass amounts of sequence data for use in research.

Alternative methods to increase the speed of which one could extract out loci in a concatenated sequence could involve aligning it to a known sequence of the locus of interest using an alignment program like BLAST (Altschul et al., 1990). However, BLAST and similar programs only align sequences that are similar, and the gene region aligned may not be entirely homologous to the gene of interest. Given that alignment programs use homologous sequences for their input, this can cause alignments that are not useful and provide the wrong phylogenetic inference, affecting downstream analyses (Lassmann & Sonnhammer, 2005).

Another major challenge to obtaining large amounts of sequence data is the highly variable nomenclature of gene names. Most genes have several alternative names and symbols that are present in sequence databases. Among distant taxa, it is common for homologous genes to vary considerably in nomenclature (Tuason et al., 2003). Even within a group of closely related taxa or within a single taxa itself, how genes are annotated may differ substantially from record to record and a wide variety of alternative gene names may be found for a single gene (Morgan et al., 2004; Fundel & Zimmer, 2006). This poses serious problems when searching through databases for data extraction like molecular sequence data (Mitchell et al., 2003; Tamames & Valencia, 2006).

Here we present the R package *AnnotationBustR* to solve the issues discussed above. *AnnotationBustR* reads GenBank annotations in R and pulls out the gene(s) of interest given a set of search terms and a vector of taxon accession numbers supplied by the user. It then writes the sequence for the gene(s) of interest to FASTA formatted files for each locus that users can then use in further analyses. For a more in depth introduction to using *AnnotationBustR* users should consult the vignette in R through `vignette("AnnotationBustR-vignette")`, which provides instructions on how to use the different functions and their respective options. Other details about the package can be accessed through the documentation via `help("AnnotationBustR")`.

Description

AnnotationBustR is written in R (R Development Core Team, 2017), a popular language for analyzing biological data. It uses the existing R packages *ape* (Paradis et al., 2004) and *seqinr* (Charif & Lobry, 2007). *AnnotationBustR* uses *seqinr*'s interface to the online ACNUC database to extract gene regions of interest from concatenated gene sequences or genomes (Gouy et al., 1985; Gouy & Delmotte, 2008). ACNUC's storage of subsequence strings allows easy access and manipulation of complex sequences, such as trans-spliced genes that may be on opposite strands of DNA. A list of the currently implemented commands is given in Table 1 and a flow chart of function usage is shown in Figure 1.

Table 1: Functions and data included in the package *AnnotationBustR*.

Function/Data Name	Description
`AnnotationBust`	Writes found subsequences for loci of interest to a FASTA file for a vector of GenBank accessions and writes a corresponding accession table.
`data(cpDNAterms)`	Loads a data frame of search terms for chloroplast genes.
`data(mtDNAterms)`	Loads a data frame of search terms for mitochondrial genes.
`data(rDNAterms)`	Loads a data frame of search terms for ribosomal DNA genes and spacers.
`FindLongestSeq`	Finds the longest sequence for each species in a set of GenBank accession numbers.
`MergeSearchTerms`	Merges two or more data frames containing search terms of features to extract into a single data frame.

The main function of *AnnotationBustR*, `AnnotationBust`, takes a vector of accession numbers and a data frame of synonym search terms to extract loci of interest and write them to a FASTA formatted file. This function also returns a pre-made accession table of all the loci of interest and the corresponding accession numbers the loci were extracted from for each species that can then be written to a csv file by the user. Users can specify duplicate genes be extracted as well, although we caution the use of doing this as they can be misleading to use in comparative analyses due to issues of paralogy (Goodman et al., 1979; Maddison, 1997). If extracting coding sequences, users can also specify if they would like to translate the sequence into the corresponding peptides by specifying the GenBank numerical translation code for the taxa of interest.

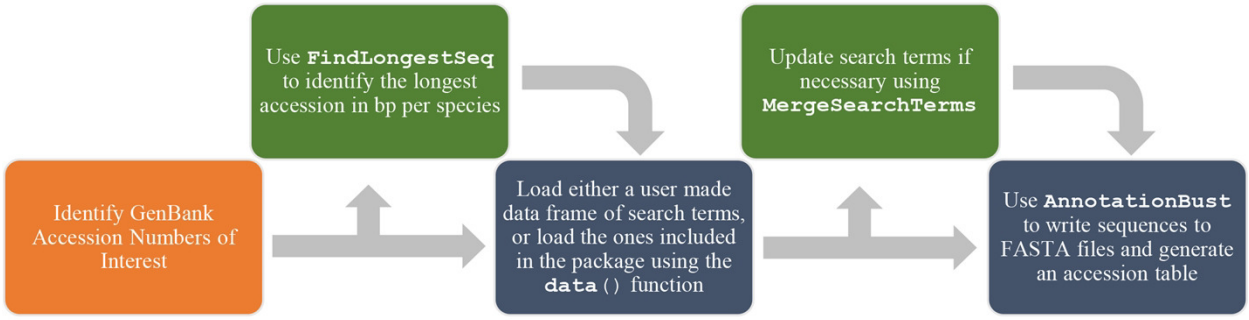

Figure 1: Flow chart of functions for a complete usage of *AnnotationBustR*. Blue boxes indicate a step using the package *AnnotationBustR* while orange boxes represent steps that need to be completed outside of *AnnotationBustR*. Boxes in green represent optional steps in the AnnotationBustR pipeline.

We have included pre-made data frames with search terms in *AnnotationBustR* for mitochondrial genomes, chloroplast genomes, and rDNA. These can be used to easily extract DNA barcodes, like cytochrome oxidase subunit I (COI) for animals in mitochondrial genomes

(Hebert et al., 2003), the internal transcribed spacers (ITS) in rDNA for fungi and plants (Kress et al., 2005; Schoch et al., 2012), and maturase K (matK) and ribulose-bisphosphate carboxylase (rbcL) genes in the chloroplast genome of plants (Hollingsworth et al., 2009). These pre-made data frames consist of three columns with the column `Locus` containing the output file name, `Type` containing the type of sequence it is (i.e. CDS, tRNA, rRNA, misc_RNA, D-loop), and the third column, `Name`, containing a possible synonym of the loci to search for. For example, for cytochrome oxidase subunit I, GenBank includes gene names of COI, CO1, COX1, cox1, COXI, cytochrome c oxidase subunit I, and COX-I. These search terms can be loaded into the workspace using the `data()` function. Annotations files for each accession are read in through *seqinr* and regular expressions matching of the synonyms provided by the user to the feature annotations are performed to identify the subsequence to extract. As certain loci may have numerous synonymous listings in GenBank feature tables that may not be included in the pre-made data frames of search terms, *AnnotationBustR* has the function `MergeSearchTerms` which allows users to easily add additional search terms to a pre-existing data frame of search terms if users follow the basic three column formatting stated above. An additional feature of *AnnotationBustR* is the function `FindLongestSeq` which finds the longest sequence for each species in a set of GenBank accessions.

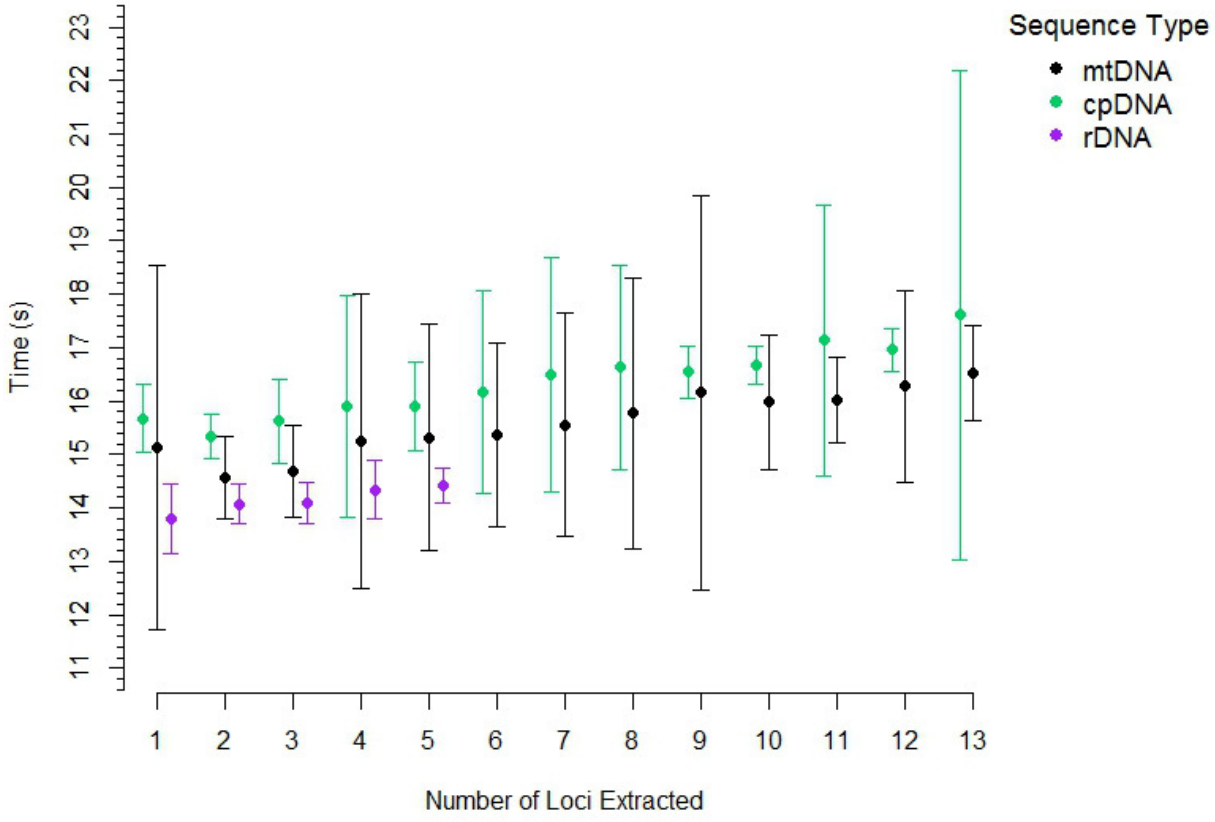

Figure 2: Timings of subsequence extraction using `AnnotationBust` for thirteen mitochondrial coding sequences (black), thirteen chloroplast subsequences (green), and five rDNA subsequence (purple). Points represent the mean time in seconds with bars representing +/- one standard deviation.

To demonstrate the performance of *AnnotationBustR*, we timed how long it took to extract thirteen popular coding sequences from 100 chloroplast genomes, the thirteen coding sequences from 100 metazoan mitochondrial genomes, and the three ribosomal RNA genes and internal transcribed spacers 1 and 2 from 100 metazoan rDNA sequences (Figure 2, see code in Supplemental Data S1). Timing trials were performed on a Windows desktop with a 3.8 GHz Intel Core i7 processor and 64 GB of RAM. For each accession, we timed the how long it took to extract one through the full number of subsequences sought. Our timings indicate that *AnnotationBustR* can efficiently extract these loci into FASTA files and that performance scales well as the number of loci to extract increases.

AnnotationBustR is available through CRAN (https://cran.r-project.org/package=AnnotationBustR) and is developed on GitHub (https://github.com/sborstein/*AnnotationBustR*). New extensions in development and fixes can be seen under the issues section on the packages GitHub page.

Conclusions

AnnotationBustR provides a quick and effortless way for users to extract subsequences from concatenated sequences or plastid and mitochondrial genomes where gene names for subsequences may vary substantially. The major limitation to the functionality of *AnnotationBustR* is that it is only as good as the annotations in the features table it is using for extraction. For instance, some concatenated sequences do not have the individual gene positions annotated for the record and just state that it contains the genes, therefore making it impossible to extract a gene from it (ex. GenBank KM260685.1, GenBank KT216295.1). Additionally, some loci may be present in the sequence yet missing from the features table completely (ex. mitochondrial D-loop missing in GenBank KU308536.1). Another limitation is that some popular loci are intergenic spacers and are not annotated in the features table, making them impossible to extract. A good example of this is the trnH-psbA intergenic spacer, a proposed locus for plant DNA barcodes (Kress et al., 2005).

Citation

Researchers publishing a paper that has used *AnnotationBustR* should cite this article and indicate the version of the package they are using. Package citation information can be obtained using citation("AnnotationBustR").

Acknowledgements

We thank members of the O'Meara lab, Cedric Landerer, and Christopher Peterson for helpful discussions while developing the package and Orlando Schwery and Frankie West for beta testing.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410. 10.1006/jmbi.1990.9999

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, and Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*:489-522.

177 Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW. 2017. GenBank.
178 *Nucleic Acids Res* 45:D37-D42. 10.1093/nar/gkw1070
179 Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *Bmc*
180 *Bioinformatics* 7:88. 10.1186/1471-2105-7-88
181 Cavender-Bares J, Kozak KH, Fine PV, and Kembel SW. 2009. The merging of community ecology and
182 phylogenetic biology. *Ecology Letters* 12:693-715.
183 Charif D, and Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical
184 computing devoted to biological sequences retrieval and analysis. *Structural approaches to*
185 *sequence evolution*: Springer, 207-232.
186 Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, and Bork P. 2006. Toward automatic
187 reconstruction of a highly resolved tree of life. *Science* 311:1283-1287.
188 De Queiroz K, and Gauthier J. 1994. Toward a phylogenetic system of biological nomenclature. *Trends in*
189 *Ecology & Evolution* 9:27-31. 10.1016/0169-5347(94)90231-3
190 Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, and Sanderson MJ. 2004. Prospects for
191 building the tree of life from large sequence databases. *Science* 306:1172-1174.
192 Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist* 125:1-15.
193 Fundel K, and Zimmer R. 2006. Gene and protein nomenclature in public databases. *Bmc Bioinformatics*
194 7:1.
195 Goodman M, Czelusniak J, Moore GW, Romero-Herrera A, and Matsuda G. 1979. Fitting the gene lineage
196 into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin
197 sequences. *Syst Biol* 28:132-163.
198 Gouy M, and Delmotte S. 2008. Remote access to ACNUC nucleotide and protein sequence databases at
199 PBIL. *Biochimie* 90:555-562. 10.1016/j.biochi.2007.07.003
200 Gouy M, Gautier C, Attimonelli M, Lanave C, and Di Paola G. 1985. ACNUC–a portable retrieval system
201 for nucleic acid sequence databases: logical and physical designs and usage. *Computer*
202 *applications in the biosciences: CABIOS* 1:167-172.
203 Hebert PD, Cywinska A, and Ball SL. 2003. Biological identifications through DNA barcodes. *Proceedings*
204 *of the Royal Society of London B: Biological Sciences* 270:313-321.
205 Hey J. 1992. Using phylogenetic trees to study speciation and extinction. *Evolution*:627-640.
206 Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW,
207 Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim K-J, Kress WJ, Schneider H, van
208 AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M,
209 Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML,
210 Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim Y-D, Lahaye R, Lee H-L, Long DG, Madriñán S,
211 Maurin O, Meusnier I, Newmaster SG, Park C-W, Percy DM, Petersen G, Richardson JE, Salazar
212 GA, Savolainen V, Seberg O, Wilkinson MJ, Yi D-K, and Little DP. 2009. A DNA barcode for land
213 plants. *Proceedings of the National Academy of Sciences* 106:12794-12797.
214 10.1073/pnas.0905845106
215 Izquierdo-Carrasco F, Cazes J, Smith SA, and Stamatakis A. 2014. PUmPER: phylogenies updated
216 perpetually. *Bioinformatics* 30:1476-1477. 10.1093/bioinformatics/btu053
217 Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, and Janzen DH. 2005. Use of DNA barcodes to identify
218 flowering plants. *Proc Natl Acad Sci U S A* 102:8369-8374. 10.1073/pnas.0503123102
219 Lassmann T, and Sonnhammer EL. 2005. Automatic assessment of alignment quality. *Nucleic acids*
220 *research* 33:7120-7128.
221 Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, and Mathews S. 2012. Hemisphere-scale
222 differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences*
223 109:16217-16221.
224 Maddison WP. 1997. Gene trees in species trees. *Syst Biol* 46:523-536.

Maddison WP. 2006. Confounding Asymmetries in Evolutionary Diversification and Character Change. *Evolution* 60:1743. 10.1554/05-666.1

Mitchell J, McCray A, and Bodenreider O. 2003. From phenotype to genotype: issues in navigating the available information resources. *Methods of information in medicine* 42:557-563.

Morgan AA, Hirschman L, Colosimo M, Yeh AS, and Colombe JB. 2004. Gene name identification and normalization using a model organism database. *Journal of biomedical informatics* 37:396-410.

O'Meara BC, Ané C, Sanderson MJ, and Wainwright PC. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60:922-933.

Paradis E, Claude J, and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.

R Development Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. *Available at* http://www.R-project.org/.

Rabosky DL, Santini F, Eastman J, Smith SA, Sidlauskas B, Chang J, and Alfaro ME. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat Commun* 4:1958. 10.1038/ncomms2958

Ratnasingham S, and Hebert PD. 2013. A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One* 8:e66213. 10.1371/journal.pone.0066213

Ratnasingham S, and Hebert PDN. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* 7:355-364. 10.1111/j.1471-8286.2006.01678.x

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Bolchacova E, Voigt K, and Crous PW. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* 109:6241-6246.

Schofl G. 2015. *reutils: Talk to the NCBI EUtils*. R package version 0.2.2.

Shi JJ, and Rabosky DL. 2015. Speciation dynamics during the global radiation of extant bats. *Evolution* 69:1528-1545.

Smith SA, Beaulieu JM, and Donoghue MJ. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol Biol* 9:37. 10.1186/1471-2148-9-37

Spriggs EL, Christin P-A, and Edwards EJ. 2014. C 4 photosynthesis promoted species diversification during the Miocene grassland expansion. *PLoS One* 9:e97722.

Tamames J, and Valencia A. 2006. The success (or not) of HUGO nomenclature. *Genome Biology* 7:1.

Tautz D, Arctander P, Minelli A, Thomas RH, and Vogler AP. 2003. A plea for DNA taxonomy. *Trends in Ecology & Evolution* 18:70-74.

Tuason O, Chen L, Liu H, Blake JA, and Friedman C. 2003. Biological nomenclatures: a source of lexical knowledge and ambiguity. Proceedings of the Pacific Symposium of Biocomputing. p 238.

Webb CO, Ackerly DD, McPeek MA, and Donoghue MJ. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*:475-505.

Winter D. 2016. *rentrez: Entrez in R*. R. 0.2.4.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, and Beaulieu JM. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506:89-92. 10.1038/nature12872